THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 10, December 1982
Printed in U.S.A.

Transmitting Data on the Phase of Speech Signals

By W. C. WONG, R. STEELE, and C. S. XYDEAS

(Manuscript received March 12, 1982)

A method for embedding data into speech signals without recourse to bandwidth expansion is proposed. Sampled speech is assembled into contiguous blocks of N samples and the Discrete Fourier Transform (DFT) is performed on each block. All the phase components in the message band, or the last J components in this band, are discarded when unvoiced or voiced speech is present, respectively. The data is introduced in the place of these rejected phase components, being $+\pi/2$ for a logical 0 and $-\pi/2$ for a logical 1. The magnitude of the coefficients associated with the data-carrying phase components are scaled to guard against data errors resulting from channel noise. The inverse DFT yields the transmitted sequence. The receiver performs the inverse process, stripping off the data and replacing it with random phase values. For an average transmission rate of approximately 1 kb/s and a channel signal-to-noise ratio of 30 dB, the bit error rate was 5.5×10^{-4}, and the average signal-to-noise ratios for voiced and unvoiced speech were 24 and −3 dB, respectively. However, the unvoiced sounds were perceived with negligible distortion owing to the preservation of their magnitude spectra. Modest error-correction codes can be used to reduce the bit error rate to 10^{-7} while maintaining the same recovered speech quality, provided the average transmitted bit rate is decreased to ≃500 b/s.

I. INTRODUCTION

Embedding data in speech signals without a significant enlargement of signal bandwidth has a great attraction if the data can be recovered without error, and the degradation of the speech is perceptually acceptable. There is a euphoria of getting a bargain, almost something for nothing. Of course it is not serendipity, but rather an exploitation of the innate redundancy in speech.

A recent proposal by Steele and Vitello[1,2] for the simultaneous

transmission of speech and data signals attempted to preserve the speech signal while accepting a bandwidth expansion of the transmitted signal. In their system the speech conveys the data using the principles of analog speech scrambling. The data becomes the scrambling key, while the receiver acts the part of a code breaker. Every time the code is deciphered correctly the receiver recovers both the data and the speech. Codes are therefore selected that are easy to break. Frequency inversion scrambling was used to achieve data rates of 700 b/s over ideal channels, and 125 b/s when additive channel noise was as high as 10 dB below the mean square value of the speech signal. In both cases there were no data errors associated with the 39,000 speech samples used in the experiments.

We now propose a system for the simultaneous transmission of speech and data that avoids a bandwidth expansion of the transmitted signal compared to that of the original speech signal, but does engender a modest reduction in the perceptual quality of the received speech.

II. THE SYSTEM

The combined transmission of data and speech in our proposed system is achieved by discarding some phase components in the speech signal and replacing them with data. At the receiver the data are removed and replaced with random phase components. By judicious choice of which phase components are used for the conveyance of data, we are able to ensure that the recovered speech quality is only marginally degraded by the presence of the data.

The speech signal bandlimited between 200 Hz and 3.2 KHz is sampled at 8 KHz and divided into sequential blocks each containing N samples. To decide whether a block of samples is to convey data, and if so, how many bits, we perform what is tantamount to a crude voice, unvoiced, or silence detection. The mean square value σ_x^2 of the samples in the block is computed and compared with two thresholds, T_1 and T_2. These thresholds float compared with the mean square value $\sum_x^2$ of the speech calculated over many blocks, such that T_1 and T_2 are α_1 and α_2 dB below $\sum_x^2$, respectively. From inspection of five sentences we experimentally determined that $\alpha_1 = 18.5$, and $\alpha_2 = 30$. The mean square value σ_x^2 is compared with these thresholds and the decision to transmit B_1 or B_2 bits of data is made according to

$$\sigma_x^2 < T_2; \text{ NO DATA TRANSMITTED} \tag{1}$$

$$T_2 \leq \sigma_x^2 < T_1; \text{ B}_1 \text{ BITS TRANSMITTED} \tag{2}$$

$$T_1 \leq \sigma_x^2; \text{ B}_2 \text{ BITS TRANSMITTED}, \tag{3}$$

where

$$B_2 < B_1. \tag{4}$$

Although our decision as to whether to transmit data, and if so, whether B_1 or B_2 bits will be embedded in the speech signal, depends only on Inequalities (1) to (3), we may consider that to a good approximation these Inequalities refer to the presence of silence, unvoiced speech, or voiced speech, respectively. Observe that as a consequence of Inequalities (1) to (3) the bit rate is variable, being dependent on the presence and nature of the speech signal. As the system is conceived for embedding data into speech signals, we envisage conventional modem techniques being deployed for the transmission of data during prolonged silences,[3,4] assuming a time assignment speech interpolation (TASI)-type arrangement is not in service.

Provided Inequality (2) or (3) is satisfied, the discrete Fourier Transform (DFT) is performed on the block of speech samples $\{x(n)\}_{n=0}^{N-1}$, namely,

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}nk}; \quad k = 0, 1, \cdots, N-1 \tag{5}$$

or

$$X(k) = Re(k) + jIm(k), \tag{6}$$

where $Re(k)$ and $Im(k)$ are the real and imaginary components of $X(k)$, respectively. The magnitude of $X(k)$ is

$$|X(k)| = \sqrt{Re^2(k) + Im^2(k)} \tag{7}$$

and its phase angle is

$$\phi(k) = \tan^{-1}\left(\frac{\mathrm{Im(k)}}{\mathrm{Re(k)}}\right). \tag{8}$$

The procedure for assigning data depends on whether Inequality (2) or (3) occurs.

2.1 Unvoiced speech

If Inequality (2) is satisfied, the speech is almost certainly unvoiced. When unvoiced speech occurs the vocal cords do not vibrate, and the sounds originate because of turbulent air flow at a constriction in the vocal tract. Unvoiced sound has a noise-like nature and tends to have low energy. The former characteristic is valuable when data, transmitted as the phase components in the unvoiced speech signal, are removed at the receiver and replaced with random phase components. The re-introduced phase components have a similar randomness to the original components, and the perceptual quality of the sound is negligibly degraded. The low energy of unvoiced speech is, by contrast, an undesirable feature when data is embedded in the phase components, as channel noise may precipitate large variations in the phase

of the received signal causing a high bit error rate (BER). Consequently, steps must be taken to increase the energy of the unvoiced sounds.

2.1.1 μ-law spectral scaling

The effect of channel noise can be mitigated by scaling the magnitude of the spectral components according to the μ-law,[5] producing magnitude components

$$|D(k)| = \frac{V \log\left\{1 + \mu_{uv} \frac{|X(k)|}{V}\right\}}{\log(1 + \mu_{uv})}; \quad |X(k)| < V$$
$$= V; \qquad |X(k)| \geq V, \tag{9}$$

where μ_{uv} is the compression factor for unvoiced speech and V is the μ-law overload parameter. The factor μ_{uv} is selected to provide an acceptably low BER, and also to contain the amplitude range of the transmitted signal. The experimental determination of μ_{uv} is discussed in Section IV. The components $|D(k)|$ are calculated for k spanning the voice bandwidth, i.e., k_{c1} to k_{c2}, where k_{c1} and k_{c2} are the spectral component associated with 200 and 3200 Hz, respectively.

2.1.2 Data insertion

Having described the scaling of the magnitude of the frequency components, we now consider how the data of B_1 bits are embedded in the phase spectrum. All the unvoiced phase components over the speech bandwidth are discarded and replaced with binary phase components determined by the data. As the phase angle $\phi(k)$ is confined to $\pm \pi$ radians, we arrange for phase components carrying data to be designated $\theta(k)$ and have values

$$\theta(k) = \pi/2, \text{ signifying logical 0}$$
$$= -\pi/2, \text{ signifying logical 1} \tag{10}$$

for $k = k_{c1}$ to k_{c2}. Although multi-level $\theta(k)$ does increase the amount of data embedded in the speech blocks, we opted for binary $\theta(k)$ to make the system more robust to channel impairments. Unless otherwise stated we will assume that every phase components contains a data bit, whence

$$B_1 = k_{c2} - k_{c1} + 1. \tag{11}$$

However, in the presence of channel impairments we may allocate each bit to an odd number of phase components, and decide on the logical value of the bit at the receiver by a simple majority vote of the logical values associated with the received phase components. More

complex channel coding techniques can be employed to further reduce BER.

2.1.3 The combined data and unvoiced speech sequence

The combined speech and data signal is obtained by performing the inverse discrete Fourier transform (IDFT) on the magnitude and phase spectral components. The original spectral coefficients are

$$X(k) = |X(k)|e^{j\phi(k)} \tag{12}$$

and those that carry data are

$$D(k) = |D(k)|e^{j\theta(k)}, \tag{13}$$

where $|D(k)|$ is given by eq. (9). The combined data and speech sequence is

$$\begin{aligned} g(i) = \frac{1}{N}\Bigg[& \sum_{k=0}^{k_{c1}-1} X(k)e^{j\frac{2\pi}{N}ik} + \sum_{k=k_{c1}}^{k_{c2}} D(k)e^{j\frac{2\pi}{N}ik} \\ & + \sum_{k=k_{c2}+1}^{N-2-k_{c2}} X(k)e^{j\frac{2\pi}{N}ik} + \sum_{k=N-1-k_{c2}}^{N-1-k_{c1}} D(k)e^{j\frac{2\pi}{N}ik} \\ & + \sum_{k=N-k_{c1}}^{N-1} X(k)e^{j\frac{2\pi}{N}ik}\Bigg]. \end{aligned} \tag{14}$$

Some values of $\theta(k)$ are used to inform the receiver whether Inequality (2) or (3) applies. This side information only constitutes a minor part of the transmitted data. The receiver is able to determine if Inequality (1) is valid by examining the mean square value of the received signal blocks.

2.2 Voiced speech

In voiced speech the vibration of the vocal cords causes broad spectrum puffs of air to excite the vocal tract, and the short-time Fourier spectrum of the speech has a quasi-periodicity, and an energy level considerably in excess of that encountered with unvoiced speech. Consequently, if data is loaded onto too many phase components, the quasi-periodicity of the recovered voiced speech will be disturbed and the speech quality degraded. We therefore discard only J phase components,

$$J < k_{c2} - k_{c1}, \tag{15}$$

whenever Inequality (3) is satisfied, and replace them with B_2 bits. Typically, J is 0.16 to 0.33 of $k_{c2} - k_{c1}$. Of course, Inequality (3) may sometimes occur when unvoiced speech is present, but only J phase components will be used for the conveyance of data. Although the

occurrence of Inequality (3) signifies that fewer phase components are available for data transmission, this state is important as voiced speech is approximately four times more prevalent than unvoiced speech. The maximum value of B_2 is J, and this will be assumed to occur unless otherwise stated. However, we note that if we assign error protection coding to the binary phase components carrying data, B_2 decreases, and so does the BER compared to the situation when B_2 is equal to J.

Spectral scaling using μ-law is also necessary for voiced speech to ensure that for the expected channel noise the BER is negligible. Equation (9) is applicable, where $|D(k)|$ is determined over the range of J coefficients. The combined voiced speech and data sequence conforms to eq. (14) with the exception that $D(k)$ extends over the range of J components.

The block diagram for embedding data into the speech signal is displayed in Fig. 1. The speech signal sampled at f_s is directed into either shift register SR_1 or SR_2, with switches S_1 and S_2 changing their positions every N/f_s seconds. While the speech samples are being extracted from SR_1, say, the computation of the mean square value σ_x^2 of the current N speech samples entering SR_2 is in progress. After N/f_s seconds σ_x^2 is determined and compared with parameters T_1 and T_2. If Inequality (1) prevails the output of comparators COMP.1 and COMP.2 are logical 0 and logical 1, respectively, and consequently the NOR gate is in the logical 0 state. Both switches S_3 and S_4 receive logical 0 signals and remain open, preventing data from being placed into the data store. When switches S_1 and S_2 change, the logical 1 state of COMP.2 is also used to inhibit the speech samples from entering the Fast Fourier Transform (FFT) circuit, and instead routes the contents of SR_2 directly to the output, bypassing the data-embedding system. This later arrangement is not shown in Fig. 1. Should Inequality (3) occur, COMP.1, COMP.2, and the NOR gate occupy logical 1, 0, and 0 states, respectively. Switch S_3 closes, and $(B_2\text{-}\epsilon)$ bits are passed into the data store, where ϵ is employed to inform the receiver that Inequality (3) applies. If both COMP.1 and COMP.2 are in the logical 0 state, the NOR gate becomes a logical 1, closing switch S_4. Data of $(B_1\text{-}\epsilon)$ bits proceed via switch S_4 into the data store, where this time ϵ signifies the presence of Inequality (2). Thus, ϵ need be only one bit, unless protection coding is added. If speech is deemed to be present, the speech in SR_2 is applied to the FFT device, and the magnitude $|X(k)|$ and phase $\phi(k)$ components of the block of speech samples generated. The $|X(k)|$ and $\phi(k)$ components are passed via switches S_5 and S_7 either directly to the Inverse Fast Fourier Transform (IFFT) via switches S_6 and S_8, or are subjected to spectral scaling and data insertion according to the number B_2 or B_1 bits removed from the data store. If voiced speech occurs only J components of $\phi(k)$ are

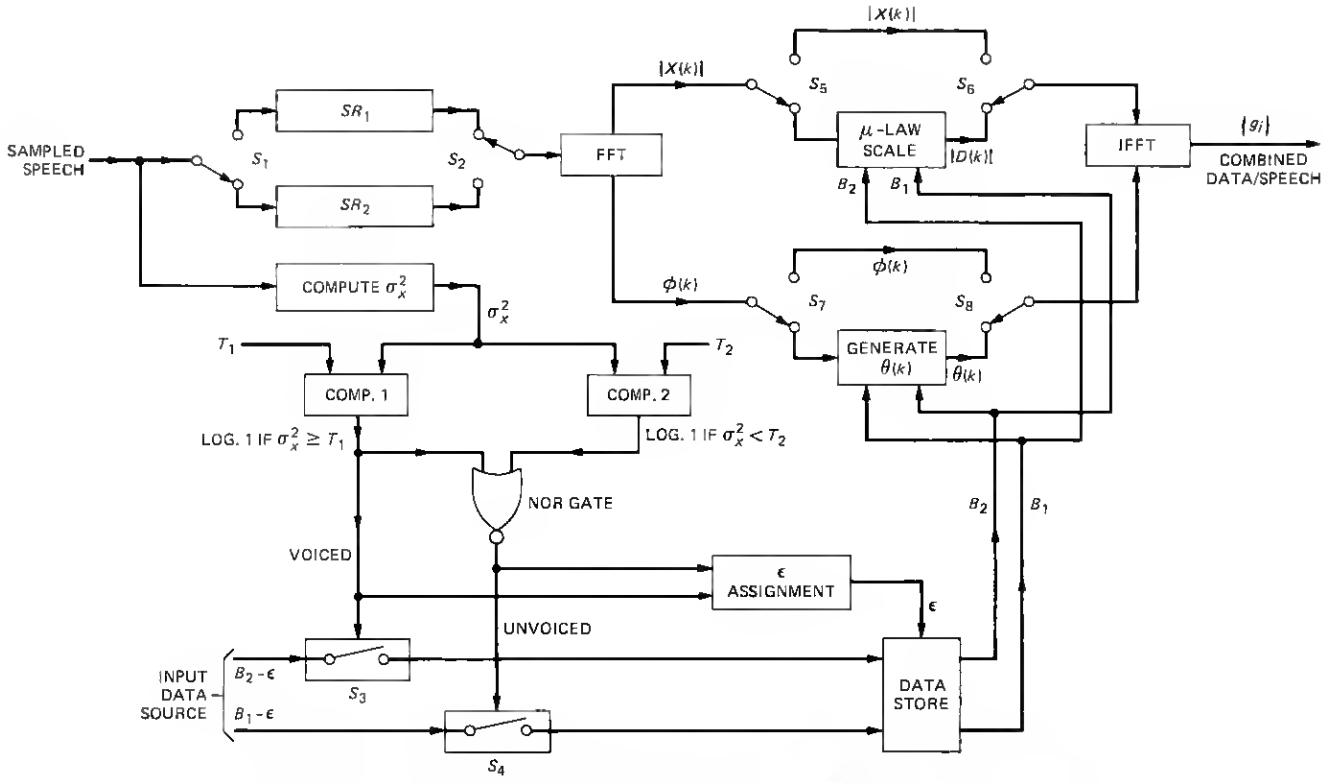

Fig. 1—Block diagram for embedding data into the phase components of the speech signal.

used, but when B_1 is present ($k_{c2} - k_{c1}$) components of $\phi(k)$ are converted to $\theta(k)$. Observe that the spectral component(s) for ϵ is always located in the same location in the J spectral region. The sequences at the output of switches S_6 and S_8 are applied to the IFFT to yield the combined speech and data sequence $\{g_i\}$.

III. THE RECEIVER

We will refrain from discussing the numerous methods by which the combined speech and data signal can be transmitted, nor will we address the variety of channels, their attendant equalization, or the techniques of correctly locking the receiver clock and the attendant acquisition of sample and block synchronization. Rather, we will assume that the combined signal is correctly sampled and ordered into the correct blocks.

The receiver's first task is to remove the data, but before that the receiver must ascertain if data have been transmitted. This is relatively straightforward since if data are embedded in the speech block, spectral scaling of the coefficients will bave been performed at the transmitter, and the mean square value of the combined signal is significantly greater than T_2 of Inequality (1). If no data is considered to be present, the received signal is accepted as the received speech signal. When data is deemed to be present, we need to determine whether it is located in J or $k_{c2} - k_{c1}$ phase components. Accordingly, the FFT is taken of the combined data and speech sequence, and the spectral phase component(s) associated with the ϵ bit(s) examined so the receiver can determine if Inequality (2) or (3) applies. Once this is accomplished the data are extracted from the received phase components $\hat{\theta}(k)$ that are known to contain binary information, according to

$$0 \leq \hat{\theta}(k) < \pi, \qquad \text{logical 0 generated}$$

and

$$-\pi \leq \hat{\theta}(k) < 0, \qquad \text{logical 1 generated.} \tag{16}$$

Having removed the data we proceed to recover the speech signal. The missing phase components are replaced by phase components having any value between $\pm\pi$ with equal probability. Those coefficients whose magnitudes were scaled are then de-scaled by inverse μ-law operation. The IFFT follows, and the speech sequence $\{\hat{x}_i\}$ so formed contains distortion, which is most serious near the ends of the blocks. A simplified explanation of this distortion is as follows. Consider successive DFT spectra to contain one line, with the phase of this line changing every block by $\pi/2$, while the amplitude of the spectral lines remains constant. The time waveform is a sinusoid whose phase

changes by $\pi/2$ at the ends of the blocks. Now consider many spectral components whose phase changes by a random value between blocks. We may again consider these components to be transformed into the time domain as sinusoids with abrupt phase changes at the block boundaries. In the case of speech, each spectral component has a different magnitude, and J or all the components may have their phases randomized. The end of block distortion ensues, its values varying from one block boundary to another in a manner difficult to quantify.

To mitigate end of block distortion we apply median filtering[6,7] to those samples at the ends of adjacent blocks. Thus, samples between the mth and $(m + 1)$th blocks are median filtered to give

$$\tilde{x}_{mN+j+i} = \underset{j=-M}{\overset{M-(2i-1)}{\text{MED}}} \{\hat{x}_{mN+j}, \hat{x}_{mN+j+1}, \cdots, \hat{x}_{mN+j+i}, \cdots, \hat{x}_{mN+j+2i}\}, \quad (17)$$

where i is a constant for a particular filter whose length is

$$L = 2i + 1, \qquad i = 1, 2, 3, \cdots, \quad (18)$$

i.e., the median value of L samples is the filtered sample. The number of samples median filtered in the vicinity of the block boundaries is

$$\lambda = 2(M + 1 - i) \quad (19)$$

and the number of samples used in the filtering of λ samples is

$$\gamma = 2(M + 1), \quad (20)$$

where M is a system parameter.

As an illustration of how the median filter equations are used, consider the example of a three-point median filter, $L = 3$, and $M = 5$. Equation (17) becomes for these parameters

$$\tilde{x}_{mN+j+1} = \underset{j=-5}{\overset{4}{\text{MED}}} \{\hat{x}_{mN+j}, \hat{x}_{mN+j+1}, \hat{x}_{mN+j+2}\}, \quad (21)$$

and $\tilde{x}_{mN+j+1}$ is the median value of $\hat{x}_{mN+j}$, $\hat{x}_{mN+j+1}$ and $\hat{x}_{mN+j+2}$. The number of samples used in the filtering process is $\gamma = 12$, which will be made up of six samples at the end of the mth block and six samples at the commencement of the $(m + 1)$th block. There are $\lambda = 10$ samples median filtered commencing with $\tilde{x}_{mN-4}$ when $j = -5$, and terminating with $\tilde{x}_{mN+5}$ when $j = 4$. Thus, the number of terms L in the brackets of Eq. (17) gives the number of samples used in the filtering process of each sample. As j steps from $-M$ to $M - (2i - 1)$, the sample being filtered, namely $\hat{x}_{mN+j+i}$, also changes under the control of j.

After λ samples have been filtered, the recovered speech sequence is obtained, having these λ samples, and $N - \lambda$ components from $\{\hat{x}_i\}$ for each block of N samples. The median filtering significantly reduces the end of block distortion.

IV. RESULTS

The sentences, "Glue the sheet to the dark blue background," "Rice is often served in round bowls," "Four hours of steady work faced us," and "The box was thrown beside the parked truck," were used in our experiments. The first two sentences were spoken by females, the remainder by males. These concatenated sentences constituting our speech signal were bandlimited between 200 Hz and 3200 Hz and sampled at 8 KHz to provide the input speech sequence, $\{x_i\}$. Random binary data were introduced into the phase components of $\{x_i\}$ in the manner described in Section II. The information ϵ was assumed to be received without error. This is a reasonable assumption because ϵ can be specified by one bit and as the error rate of the phase components carrying data will be shown to be 0.055 percent, the probability of ϵ being in error can be rendered negligible by assigning a small number of error-correcting bits to ϵ. The block size N was 256.

Because of the spectral scaling, tbe first experiments related to the increases in the peak and rms values of the combined speech and data sequence, $\{g_i\}$, compared to those of the input speech sequence, $\{x_i\}$. We were concerned that the spectral scaling might significantly increase the amplitude and power levels of the original speech signal and overload the communication channel. To observe the effect of spectral scaling on the amplitude components in $\{g_i\}$ we proceeded as follows. Each block of speech was examined, and those blocks where Inequality (3) applied, our so-called voiced blocks, were noted. Using these blocks we calculated two signal expansion parameters, which we defined as,

$$r_v \triangleq \frac{1}{\psi_v} \sum_{i=1}^{\psi_v} \frac{|g|_{\max,v,i}}{|x|_{\max,v,i}} \tag{22}$$

and

$$\rho_v \triangleq \frac{1}{\psi_v} \sum_{i=1}^{\psi_v} \frac{\Omega_{g,v,i}}{\Omega_{x,v,i}}, \tag{23}$$

where $|g|_{\max,v,i}$ and $|x|_{\max,v,i}$, and $\Omega_{g,v,i}$ and $\Omega_{x,v,i}$ are the maximum and rms values of the combined speech and data sequence, and tbe input speech sequence, in the ith blocks, respectively. The number of voiced blocks is ψ_v, where the subscript v is used to signify the applicability of Inequality (3). Figures 2 and 3 show the variation of r_v and ρ_v as a function of the spectral scaling factor, μ_v, for voiced speech. As more phase components are used to convey data, i.e., increasing J, more spectral magnitude components are increased by μ-law spectral scaling; and on performing the IFFT the rms and maximum amplitudes of the voiced blocks in the transmitted signal are increased, and hence r_v and ρ_v are increased for a given μ_v. Similarly, for a given J the effect of

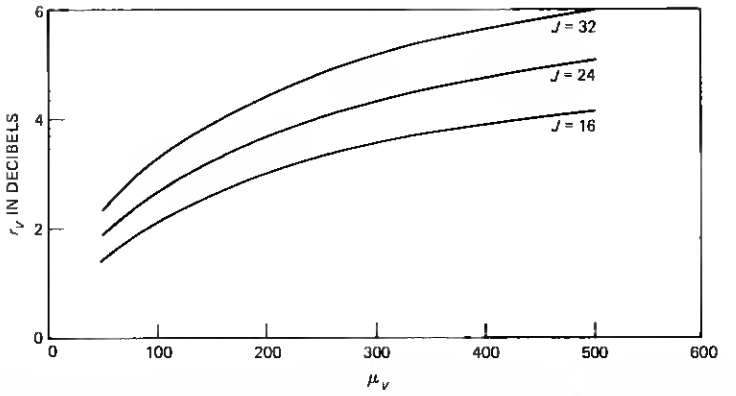

Fig. 2—Variation of the average ratio of maximum amplitudes of the transmitted signal to the input speech signal, r_v, as a function of the μ-law scaling factor, u_v, for voiced speech for different values of J.

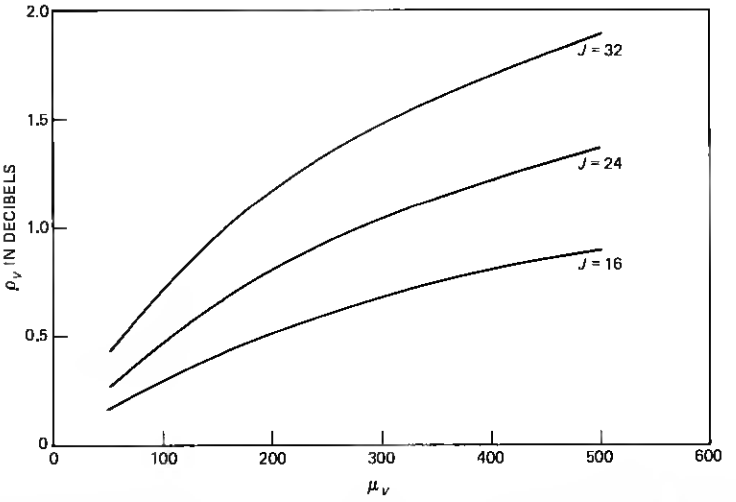

Fig. 3—Variation of the average ratio of the rms value of the transmitted signal to the input speech signal, ρ_v, as a function of the μ-law scaling factor u_v for voiced speech for different values of J.

increasing μ_v is to increase the spectral scaling of the magnitude components, which consequently increases r_v and ρ_v.

By selecting only those blocks where Inequality (2) applied, we found the signal expansion factors r_{uv} and ρ_{uv} using the same procedures as employed for r_v and ρ_v [see eqs. (22) and (23)]. The subscript uv, an abbreviation for unvoiced speech, implies the validity of Inequality (2). The variation of r_{uv} and ρ_{uv} as a function of the spectral scaling factor, μ_{uv}, for unvoiced speech is displayed in Fig. 4. Unlike

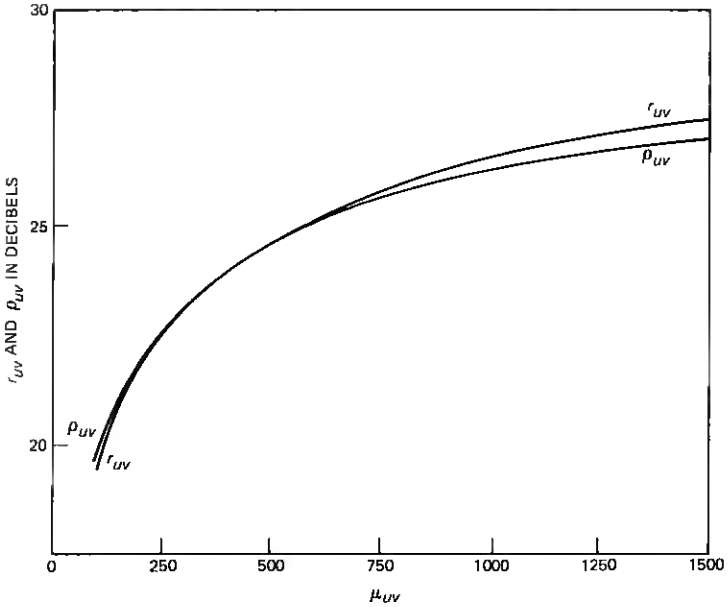

Fig. 4—Variation of r_{uv} and ρ_{uv} as a function of μ_{uv}.

the small increases in r_v and ρ_v that occur for voiced speech, the effect of spectral scaling all the magnitude components by large values of μ_{uv} results in substantial increases in r_{uv} and ρ_{uv}. However, unvoiced speech has much lower magnitude and rms values than voiced speech, enabling much larger values of r_{uv} and ρ_{uv} to be used, thereby protecting the data from channel noise. If it is required that the peak or rms value of $\{g_i\}$ is not to exceed that of $\{x_i\}$, an attenuator must be placed after the IFFT in Fig. 1. The effect of such an attenuator on the recovered signal-to-noise ratio (s/n) and BER will be discussed later.

An objective criterion for the quality of the recovered speech signal should take cognizance of the particular process being used to convey data, yet be sufficiently well known to have comparative value. In this system the distortion in the recovered speech signal originates from two main processes, namely, the randomizing of the phase components of voiced and unvoiced speech, and the effect of additive channel noise. The randomization of the phase components does not alter the recovered magnitude spectra, and thus spectral distortion measures[8] based on spectral power are inappropriate. The only errors in the magnitude spectra derive from the channel noise. Signal-to-noise ratio measurements are familiar to engineers in spite of their shortcomings, and the two most widely quoted are the average s/n and the segmental s/n.[9] In the former the ratio of the average signal power to the average error power is found. In determining segmental s/n the signal is divided

into segments or blocks, and the average signal to average error power is computed in decibels for each segment. Then s/n values of each segment are averaged to give segmental s/n. We decided to use segmental s/n, and proceeded to divide the speech into voiced or unvoiced segments. We made this division because the effect of randomizing the phase yields s/n values for a segment that critically depend on whether the segment contains voiced or unvoiced speech. A low s/n for unvoiced speech can be anticipated, as randomizing every phase component yields a time waveform that is radically different from the original segment of speech. The s/n for that block of speech is accordingly very low, and is often negative. However, because the magnitude of the spectra for the recovered and original speech signals are the same, these signals are perceived to be similar. In the case of voiced speech the effect of randomizing J spectral components without altering their magnitudes results in end of block distortion. This distortion is mitigated by employing median filtering as previously described. The end of block distortion is not significant with unvoiced speech because of the relatively small magnitudes of the spectral coefficients. By measuring the s/n of each voiced segment, we provide a measure of the end of block distortion. Thus, segmental s/n is a reasonable measure for voiced speech, and a poor measure for unvoiced speech, in that the value of the segmental s/n has a close correspondence with the perceived speech in the case of voiced speech, and vice versa for unvoiced speech. We note in passing that in waveform encoding, like the situation here, the segmental s/n is usually high for voiced speech and low or negative for unvoiced speech.[10]

In our experiments we proceeded as follows. Assuming the channel to be ideal we determined the s/n of the recovered speech signal as a function of the number J of phase components discarded for voiced speech. Only blocks where Inequality (3) applied were used in the s/n calculation. We performed experiments for J measured over different coefficient ranges, e.g., from k_{c1} to higher values of k, about the center of the coefficient range, and from k_{c2} to lower values of k. The location of the range of J caused different perceptual impairments in the recovered speech signal. From informal listening tests we concluded that the latter range for J was preferable, and, accordingly, we display in Fig. 5 the s/n for voiced speech as a function of J measured from k_{c2} to lower values of k. In determining the s/n, we employed the median filter having a length L of 3, and $M = 5$. As the curve in Fig. 5 was obtained for an ideal channel, it is independent of the values of μ_v and μ_{uv}, factors introduced to avoid data errors in the presence of channel impairments. The exchange of s/n in decibels with J is given by

$$\text{s/n} \simeq 39 - 0.375J; \qquad 8 \leq J \leq 40, \tag{24}$$

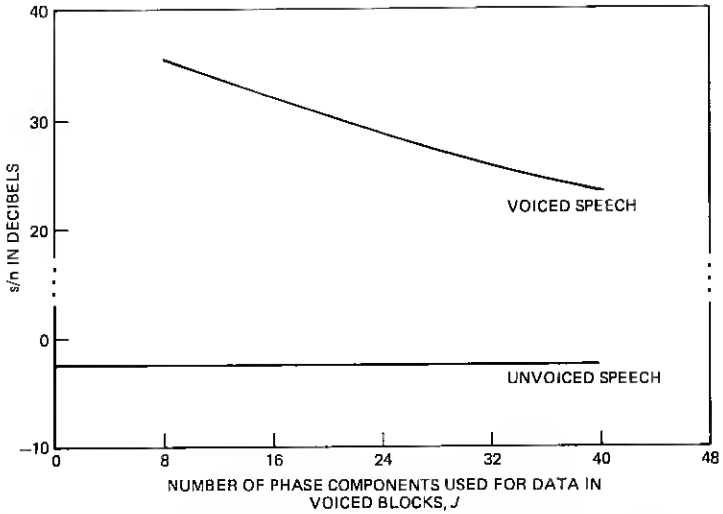

Fig. 5—Variation of s/n versus J for voiced speech. The s/n of unvoiced speech is also shown.

i.e., a loss of 0.375 dB in s/n per phase component of voiced speech employed for the transmission of data.

In the case of blocks containing unvoiced speech 98 phase components in each block are used for the conveyance of data (corresponding to 200 to 3200 Hz, $N = 256$), and the segmental s/n for these blocks is only −2 dB. As mentioned above, this low s/n for unvoiced speech was expected owing to the randomization of all the phase components in the recovered speech. However, the perceptual quality and intelligibility of the recovered unvoiced speech is good as the magnitude spectra are maintained, and the excitation in unvoiced speech is noise-like. The s/n for unvoiced speech is shown in Fig. 5 as a horizontal line.

For the sentences used, transmitted data rates of 1200, 1024, and 852 b/s were achieved for $J = 32$, 24, and 16, respectively.

To prevent the channel from being overloaded by excessive amplitude levels resulting from signal amplification owing to spectral component scaling, attenuation of $\{g_i\}$ was performed. The attenuation was adjusted until the range of amplitude levels of the combined speech and data signal was the same as that of the input speech signal. Specifically, we found the block with the largest output amplitude whose magnitude expansion parameter was r_k, say. The attenuation in decibels was then set at

$$I = 20 \log_{10}(r_k)$$

for the whole speech signal. Channel noise $\{n_i\}$ was next added to the transmitted signal, and for a constant channel noise power of minus P

dB below the mean square value of the input speech signal, the change in s/n relative to the s/n in the absence of spectral scaling was found as a function of μ_v for blocks where Inequality (3) applied. This was repeated for different values of P and two values of J to yield the curves shown in Fig. 6a and b. As we expected, when P becomes progressively more negative, the change in s/n, namely Δs/n, approaches zero for all μ_v. When the additive noise power P is high and $J = 32$, there is a loss in s/n that increases with μ_v, but never exceeds 3 dB for the parameters shown in Fig. 6a. For $J = 24$, the loss in s/n is much smaller ($\simeq$1 dB), and for μ_v below 100, Δs/n may be slightly positive. This small positive value of Δs/n arises because for low values of μ_v, the spectral scaling of the J coefficients carrying data is insufficient to cause the combined data and speech sequence $\{g_i\}$ to be attenuated. Thus, for a given channel noise power P the channel s/n decreases with tbe result that the recovered s/n is marginally enhanced. When μ_v exceeds 100, and the attenuation of $\{g_i\}$ is as described above, the channel s/n decreases, and Δs/n takes on negative values. When the experiment was repeated with blocks where Inequality (2) applied, Δs/n was always positive as shown in Fig. 6c. Observe that for unvoiced speech no attenuation of the combined speech and data signal need be imposed for $\mu_{uv} \leq 400$, as the signal does not exceed the levels found in voiced speech. Consequently, Δs/n is nearly constant until $\mu_{uv} > 400$, whence attenuation of $\{g_k\}$ is employed. Δs/n decreases slowly with μ_{uv}, and Δs/n is marginally greater for P of -20 dB than -30 dB, i.e., -20 dB of channel noise is advantageous. However, the variation of Δs/n in Fig. 6 is not great, being positive for unvoiced speech and negative (in general) for voiced speech.

With the attenuator adjusted as previously described such that the amplitude range of the transmitted signal and the original speech signal are the same, we observed an improvement in BER, defined as

$$IBER = 20 \log_{10}\left\{\frac{BER_\mu}{BER_o}\right\}, \tag{25}$$

where BER_μ and BER_o represent the BER when spectral scaling of value μ is used and when no spectral scaling is employed, respectively. The variation of IBER with μ for different values of noise power P is displayed in Fig. 7. The smallest value of P employed was -30 dB, as we did not have sufficient data for reliable results when P was more negative. We see from Figs. 7a and b that μ_v of the order of 250 is a good choice as it provides a large value of IBER while avoiding significant losses in s/n, as displayed in Figs. 6a and b. Thus, for $J = 24$, $P = -30$ dB, a $\mu_v = 250$ provides a gain in BER of $\simeq$50 dB while sustaining a loss in recovered speech s/n of only 0.5 dB. As is expected,

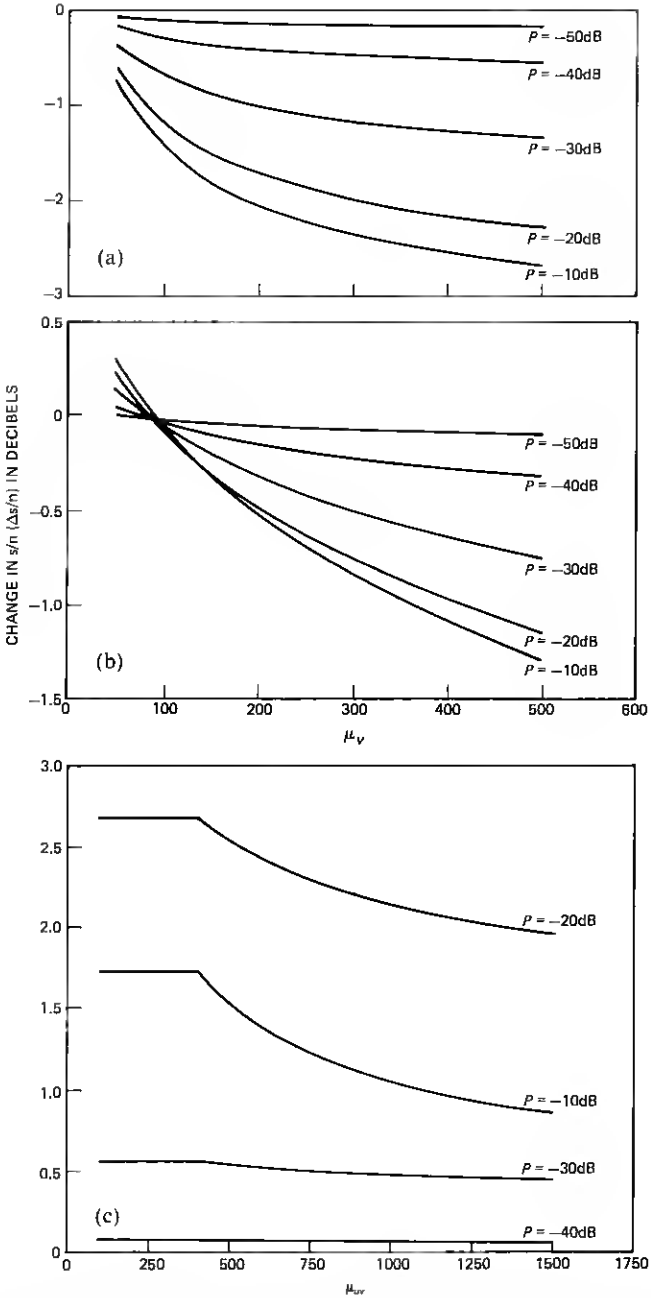

Fig. 6—Effect of spectral scaling on s/n. The change (Δs/n) in s/n relative to the s/n in the absence of spectral scaling, for different channel noise power P. (a) Δs/n versus μ_v, $J = 32$. (b) Δs/n versus μ_v; $J = 24$. (c) Δs/n versus μ_{uv}.

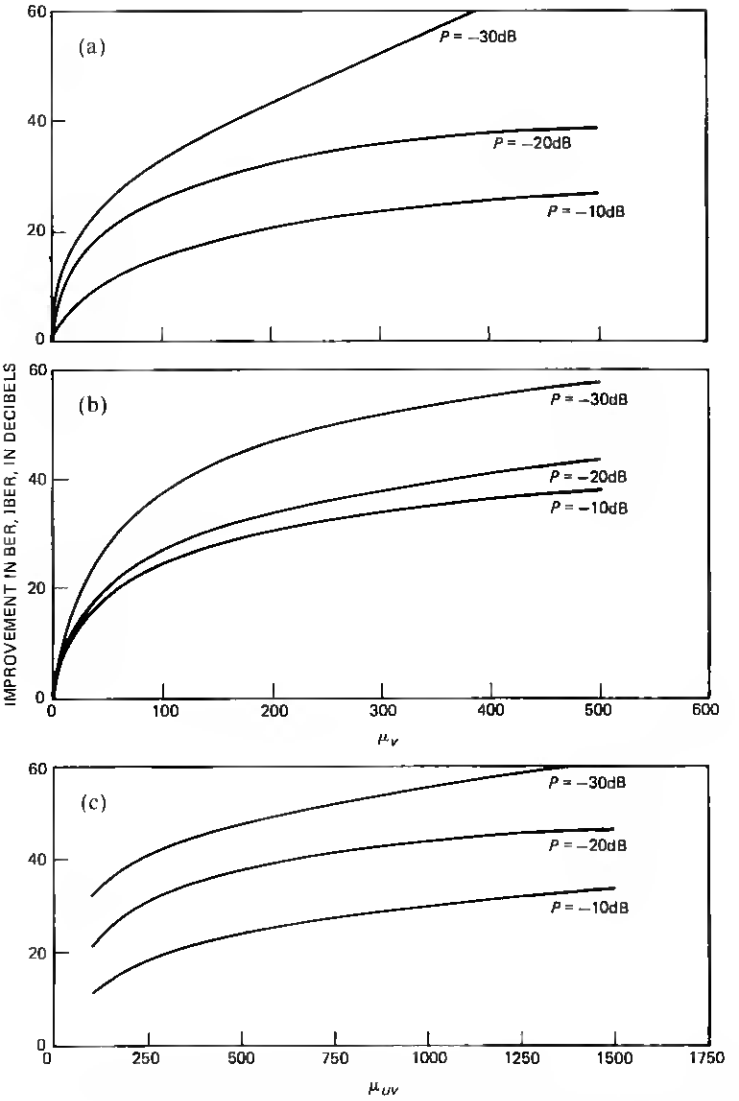

Fig. 7—Improvement in BER, namely IBER, due to spectral scaling, for different channel noise power P. (a) IBER versus μ_v, $J = 32$. (b) IBER versus μ_v, $J = 24$. (c) IBER versus μ_{uv}.

larger values of μ_{uv} apply as shown in Fig. 7c, and a good choice of μ_{uv} is 750. Using this μ_{uv} value, $P = -30$ dB, we achieved IBER of ≃50 dB and a gain in s/n of 0.5 dB. Figures 6 and 7 highlight the desirable properties of spectral scaling, a large improvement in BER, and at worst a small loss in speech s/n.

The channel s/n was computed as

$$s/n_c = 10 \log_{10} \left\{ \frac{\sum_{i=1}^{W} \tilde{g}_i^2}{\sum_{i=1}^{W} n_i^2} \right\}, \tag{26}$$

where $\tilde{g}$ is the attenuated version of g, and W is the number of speech samples in the input speech signal. Although the same noise source was used as in the previous experiments, and the attenuator was employed, s/n_c differs from the s/n of P dB computed using the input speech and the noise signal. This difference arises because $\{\tilde{g}_i\}$ is not identical to $\{x_i\}$. The sequence $\{\tilde{g}_i\}$ depends on μ_v and μ_{uv}, parameters which affect both r and ρ [see eqs. (22) and (23)]. However, the s/n differences are small, and typically are <3 dB. Figure 8 displays the

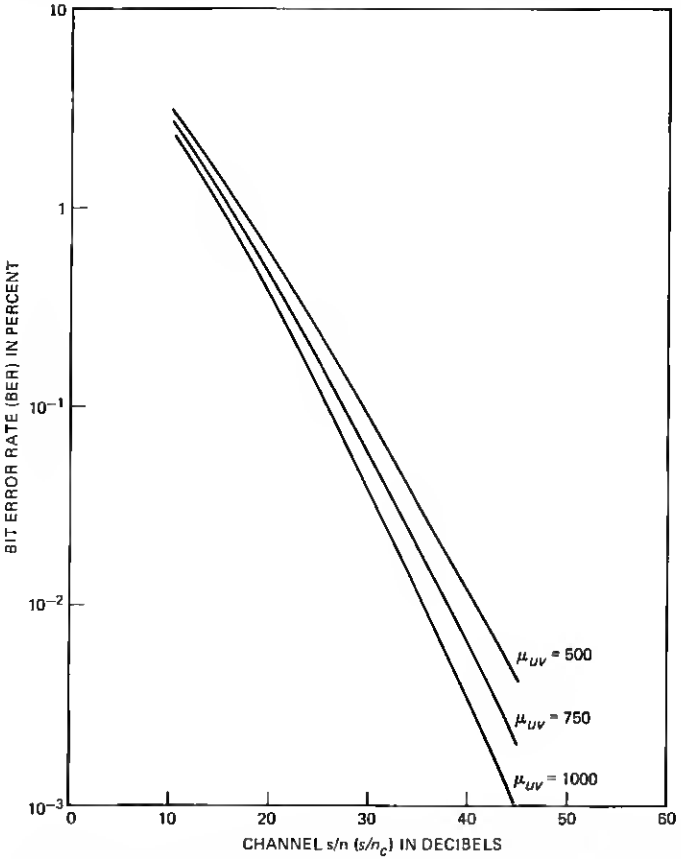

Fig. 8—Variation of BER as a function of s/n_c for different values of μ_{uv}; $\mu_v = 250$, $J = 24$.

variation of BER as a function of s/n_c for different μ_{uv}, and $\mu_v = 250$. As we anticipated from Fig. 7c, increasing μ_{uv} from 500 to 1000 results in an increase in IBER, which means a decrease in BER. Further increases in μ_{uv} will decrease BER, but the reduction will not be great.

We observe from Fig. 8 that for s/n_c of 30 dB, the BER is 0.055 percent, and as previously stated, the average transmitted bit rate for $J = 24$ is 1024 b/s. This BER can be reduced by using error-correcting codes. For example, if a BCH code is employed such that the number of error-correcting code bits equals the number of data bits, i.e., the average transmission rate is 512 b/s, the BER decreases to approximately 10^{-5} percent. The extent of the trading of the reduction in the average transmitted bit rate for improvements in BER depends on system requirements. In digital radio transmission outage occurs when the BER exceeds 0.01 percent.

The variation of the segmental s/n of the recovered speech signal against s/n_c is shown in Fig. 9 for three values of J, and $\mu_v = 250$ and $\mu_{uv} = 750$. Only blocks satisfying Inequalities (2) and (3) were used in this calculation of segmental s/n. As s/n_c approaches 50 dB we approximate to the ideal channel condition, and by comparing the s/n values with those in Fig. 5 we may observe the deleterious effect of the unvoiced speech s/n on the overall s/n. Thus, for $J = 16$, 24, and 32, the s/n = 25, 23, and 20.5 dB in Fig. 9, whereas when only voiced speech is present the corresponding s/n = 32, 28, and 26 dB. However, the perceptual quality of the recovered speech is more suitably represented by the segmental s/n for voiced speech than by the combined segmental s/n. Thus, the s/n values in Fig. 9 are lower than would be anticipated for the quality obtained. If no phase components had been used for the transmission of data, the variation of s/n with s/n_c would be a straight line at 45 degrees, shown in Fig. 9. The offset of this line from the origin is due to the s/n of the speech being calculated as segmental s/n,[9] and s/n_c is computed according to eq. (26).

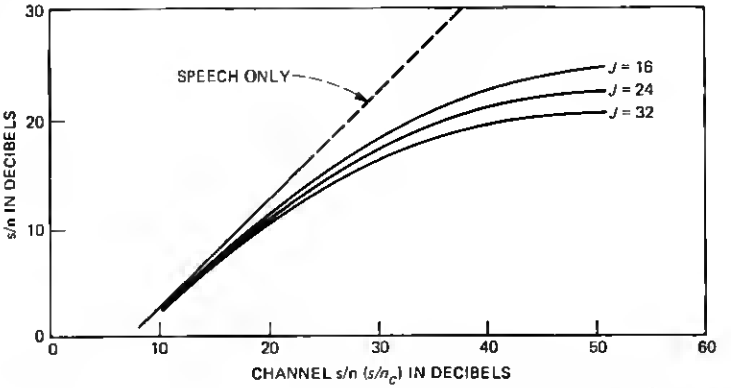

Fig. 9—Recovered s/n as a function of s/n_c for $J = 16$, 24, and 32, $\mu_v = 250$, $\mu_{uv} = 750$.

We conclude this section by providing some waveforms and spectra of the input, transmitted, and recovered signals for $\mu_v = 250$, $\mu_{uv} = 750$, $J = 24$, and $N = 256$. In Fig. 10a twenty blocks of an arbitrary speech signal are shown, having two blocks of intraconversational silence. The combined speech and data sequence $\{g_i\}$ prior to attenuation is displayed in Fig. 10b, where it can be observed that the power level of the unvoiced speech is considerably amplified; high-amplitude, high-frequency components have been introduced into the voiced segments; and those parts of the silence that resided in blocks substantially occupied by voiced speech are carrying data. The recovered speech signal is displayed in Fig. 10c for the case of an ideal channel. The effect of replacing the data-carrying phase components by random ones does not cause serious degradations in the perceptual quality of the recovered speech.

The magnitude of the spectral components of the waveforms in Figs. 10a and b are shown in Figs. 11a and b, respectively. As data is carried by the phase components in the speech signal, the magnitude spectra of the waveforms in Fig. 10a and c are identical. The μ-law scaling of 24 components for voiced speech is seen to substantially enhance its high-frequency components, whereas all 98 components across the speech band are scaled for the unvoiced speech. The μ-law scaling for voiced speech is seen to be reminiscent of frequency pre-emphasis.

V. DISCUSSION

A system has been proposed for the simultaneous transmission of speech and data on the phase of the speech signals, where the bandwidth of the transmitted signal is contained relative to that of the original speech. We knew at the outset that if data was to be conveyed on the phase of speech signals, the receiver would be forced to introduce phase components to replace those that had been discarded at the transmitter in favor of data. We postulated that if the introduced phase components were derived from a random number source, and that their values were confined between $\pm\pi$, then the perceptual degradation in speech quality might be acceptable. Our decision to randomize the values of the introduced phase components at the receiver was based on the knowledge that the variations in the values of the phase spectral components in speech, particularly unvoiced speech, tend to have random behavior. Further, the effect of phase distortion on monaural speech intelligibility is known to be small, the controlling factor being the amplitude spectra. Accordingly, we did experiments, and from informal listening experiences concluded that the randomization of all the phase components of unvoiced speech did not cause serious perceptual degradation. In the case of voiced speech we discovered that if too many phase components were randomized

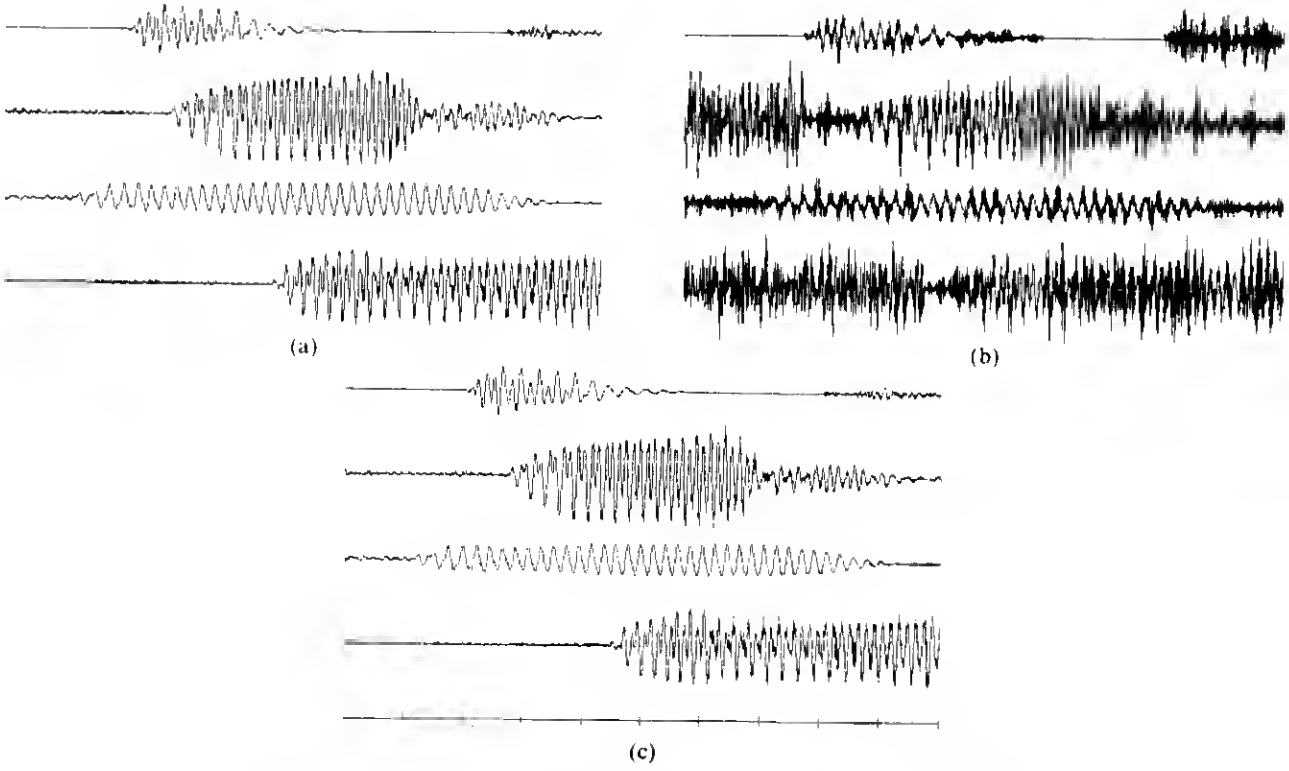

Fig. 10—(a) Arbitrary speech waveform. (b) Combined speech and data signal prior to attenuation. (c) Recovered speech signal; transmission via an ideal channel.

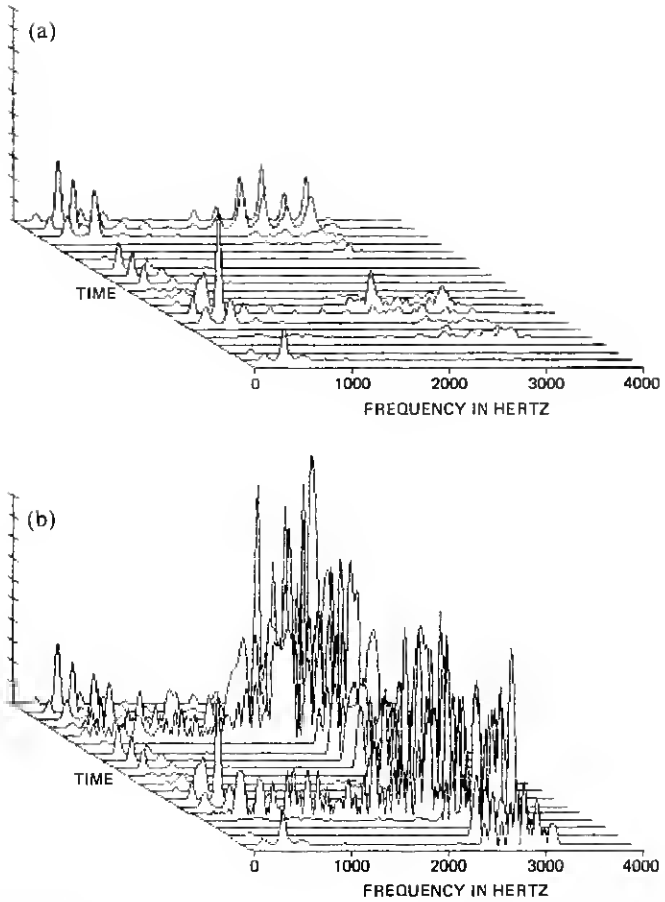

Fig. 11—Magnitude spectra. (a) The signal in Fig. 10a and c. (b) The signal in Fig. 10b.

the recovered speech quality was poor, and out of the 98 phase components available in our experiments we concluded that the maximum number J of phase components that could be randomly perturbed was 32. The position of the J components had different perceptual effects, and we decided to make J span the range of the highest inband frequency components, although the actual position of J is far less important than its value.

Deeming that randomization of the phase components as described was perceptually tolerable, particularly in the presence of channel noise when the channel s/n was approximately 30 dB, we decided to trade the loss in perceptual quality for the implantation of data into those components we had randomized. By this strategy, and for a channel s/n of 30 dB we have been able to achieve an average data

rate of 1 Kb/s on the assumption that one bit is assigned to eacb data-carrying phase component. To achieve this data rate we are required to tolerate a BER of 0.055 percent and an average s/n for voiced and unvoiced speech of 24 and −3 dB, respectively, the measurements being made over four sentences of speech. Observe that the BER can be reduced to a value acceptable for the user by applying channel-coding strategies that result in a reduction in the transmitted bit rate. An example of such a trade-off is given in Section IV. The recovered speech is below toll quality, but the ability to transmit data may make this quality reduction acceptable in certain situations.

Making comparisons of this technique of conveying data on the phase of the speech signal with those employing scrambling methods[1,2] is difficult because of the radically different approaches of these schemes. Embedding data in speech by scrambling can be made to have a very small bandwidth expansion by suitable choice of scrambling code and block size.[2] Increasing the complexity of the scrambling algorithm and the number of bits per block of speech scrambled alters the systems performance in a way that is difficult to predict.

The previously described system using scrambling techniques,[1,2] and the one described here have only been evaluated for noisy channels. Which system would perform best in an actual communications network, and what the requirements would be on channel equalization and synchronization are unknown quantities. What we can say is that errors in the samples at the receiver attributable to noise or imperfect channel equalization, the presence of an unwanted sample, and the loss of a wanted one owing to incorrect synchronization, are smeared over the spectral components by the DFT. The data here is binary and therefore considerable noise on the data-carrying phase can be tolerated. By using error detection and correction coding the data rate can be sacrificed to a value commensurate with a specified BER for a given set of channel impairments.

However, our quest was not to investigate the numerous channel conditions. It was to determine if speech and data could be transmitted over a noisy channel by embedding the data in the phase of speech, and further, if the transmitted bit rate could be sufficiently high to be useful, the BER acceptably low, and the degradation in the recovered speech quality perceptible but not annoying. Our conclusion is affirmative.

REFERENCES

1. R. Steele and D. Vitello, "Simultaneous Transmission of Speech and Data Using Code-Breaking Techniques," B.S.T.J., *60*, No. 9 (November 1981), pp. 2081–2105.
2. R. Steele and D. Vitello, "Embedding Data in Speech Using Scrambling Techniques," ICASSP '82, Paris, *3* (May 1982), pp. 1801–4.
3. P. T. Brady, "A Statistical Analysis of On-Off Patterns in 16 Conversations," B.S.T.J., *47*, No. 1 (January 1968), pp. 73–91.

4. M. R. Schroeder and S. L. Hanauer, "Interpolation of Data with Continuous Speech Signals," B.S.T.J., *46*, No. 8 (October 1967), pp. 1931–3.
5. B. Smith, "Instantaneous Companding of Quantized Signals," B.S.T.J., *36*, No. 3 (May 1957), pp. 653–709.
6. J. W. Tukey, "Nonlinear (Nonsuperposable) Methods for Smoothing Data," EASCON Record (1974), p. 673.
7. R. Steele and D. J. Goodman, "Detection and Selective Smoothing of Transmission Errors in Linear PCM," B.S.T.J., *56*, No. 3 (March 1977), pp. 399–409.
8. J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere, "A Comparison of the Performance of Four Low-Bit-Rate Speech Waveform Coders," B.S.T.J., *58*, No. 3 (March 1979), pp. 699–712.
9. D. J. Goodman, C. Scagliola, R. E. Crochiere, L. R. Rabiner, and J. Goodman, "Objective and Subjective Performance of Tandem Connections of Waveform Coders with an LPC Vocoder," B.S.T.J., *58*, No. 3 (March 1979), pp. 601–29.
10. C. S. Xydeas, C. C. Evci, and R. Steele, "Sequential Adaptive Predictors for ADPCM Speech Encoders," IEEE Trans. Commun., *COM-30*, No. 8 (August 1982), pp. 1942–54.